IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 16, 2018, accepted October 12, 2018, date of publication November 19, 2018,
date of current version December 27, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2882227

Integrated Deep Model for Face Detection and Landmark Localization From "In The Wild" Images

GARY STOREY, AHMED BOURIDANE, AND RICHARD JIANG
Department of Computer and Information Sciences, Northumbria University, Newcastle Upon Tyne NE1 8ST, U.K.
Corresponding author: Richard Jiang (richard.jiang@northumbria.ac.uk)

This work was supported by the EPSRC under Grant EP/P009727/1.

ABSTRACT The tasks of face detection and landmark localisation are a key foundation for many facial analysis applications, while great advancements have been achieved in recent years there are still challenges to increase the precision of face detection. Within this paper, we present our novel method the Integrated Deep Model (IDM), fusing two state-of-the-art deep learning architectures, namely, Faster R-CNN and a stacked hourglass for improved face detection precision and accurate landmark localisation. Integration is achieved through the application of a novel optimisation function and is shown in experimental evaluation to increase accuracy of face detection specifically precision by reducing false positive detection's by an average of 62%. Our proposed IDM method is evaluated on the Annotated Faces In-The-Wild, Annotated Facial Landmarks In The Wild and the Face Detection Dataset and Benchmark face detection test sets and shows a high level of recall and precision when compared with previously proposed methods. Landmark localisation is evaluated on the Annotated Faces In-The-Wild and 300-W test sets, this specifically focuses on localisation accuracy from detected face bounding boxes when compared with baseline evaluations using ground truth bounding boxes. Our findings highlight only a small 0.005% maximum increase in error which is more profound for the subset of facial landmarks which border the face.

INDEX TERMS Computer vision, face detection, machine learning.

I. INTRODUCTION

The task of object detection is a foundation computer vision problem which has seen a number of advances in recent years through the application of deep learning methods. Convolutional layer architectures such as AlexNet [1], VGG [2] and ResNets [3] have provided precise feature learning methods for reliable object detection, while object proposal methods such as Faster R-CNN [4] and YOLO9000 [5] have significantly improved the computational efficiency for detecting objects at multiple scales. Facial detection is a vastly popular object detection task due to the many application domains associated with the human face such as facial recognition [6] and facial expression analysis [7], [8]. Although face detection is well researched challenges still remain, most specifically in unconstrained ''In The Wild'' images where extreme poses, occlusion, tiny faces, resolution variations and de-focus have a large effect on facial appearance, while there is also a challenge to increase the precision of methods through the reduction of false positive detection's. Traditionally the face detection task was tackled singularly but in recent years there has been some success applying multi-task methods that also incorporate landmark localisation and pose estimation tasks [9]–[11].

The computer vision tasks of face detection and landmark localisation have a rich research history with many methods being proposed. Major breakthroughs for the face detection task include scanning window classifiers [12] which applied a cascaded approach for real-time detection, while the Deformable Part Model (DPM) [9], [13] based techniques increased accuracy and tackled multiple viewpoints but with increased computational overhead. Recently exceptional results have been achieved by methods applying deep Convolutional Neural Networks (CNN) [14]–[16] eclipsing those of the traditional methods. Landmark localisation also referred to as face alignment is the process of locating semantically meaningful facial landmarks which compose facial components such as brows, eyes, lips and nose. Research to date can be generally divided into three categories. Holistic based approaches such as Active Appearance Models (AAM's) [17], [18] solve the face alignment problem by jointly modelling appearance and shape. Local expert based methods such as Constrained Local Model (CLM's) [19]

FIGURE 1. Examples of Integrated deep model outputs from an "In The Wild" image.

learn a set of local experts detectors or regressors [20], [21] and apply shape models to constrain these. The most recent advancements which have attained state-of-the-art results apply CNN based architectures with probabilistic landmark predictions in the form of heat maps as the network output [22].

Research has identified that combining the traditionally individual tasks of face detection and landmark localisation into unified methods can boost the accuracy of both tasks [9], [11]. The methods employed include feature fusion at different levels of a network, singular shared feature sets learnt by joint optimisation functions representing each of the tasks [4] or both [11]. Performance increases are attributed to the inter-connectivity of the tasks. The original unified model combining face detection, pose estimation and alignment was the multi-view Trees Structured Model (TSM) [9], at the time of publication this approach advanced the state-of-the-art for each of the associated tasks. One specific drawback of the TSM method [9] is that there is a high level of trade off between accuracy and computationally overhead.

In this paper we present a novel method which we term Integrated Deep Model (IDM) for the joint tasks of face detection and landmark localisation from unconstrained "In The Wild" images. Our method integrates two state-of-the-art deep convolutional networks, the first is the Faster R-CNN model [4] and the second a stacked hourglass model [22]. The aim of the proposed IDM method is to leverage the inter-connectivity of the learnt features in both architectures to increase the precision of face detection while also providing accurate landmark localisation. An example output from the proposed IDM is given in Fig. 1.

To summarise our main contributions in this paper are as follows:

(1) We propose the Integrated Deep Model with the aims to leverage the strengths of both architectures to improve face detection precision, through the introduction of a joint learning optimisation function.

(2) We propose a transformation method for the heat map output of the stacked hourglass model so it can be applied to the task of face detection in addition to the current landmark localisation output.

(3) Through experimental results we investigate the impact integration has on the accuracy of both face detection and landmark localisation.

The remainder of this paper comprises of a review of relevant work within section II, followed by an in-depth overview of the IDM method within section III. Section IV is a discussion of the experiments undertook and the results gained and a conclusion is given within section V.

II. RELATED WORK

Within computer vision there has been a long history of approaches to both face detection and face landmark localisation tasks. Initially approached as separate tasks these have more recently been researched as integrated methods. The focus of this section is to provide an overview of previously proposed methods as space does not allow for a full review.

A. FACE DETECTION

Face detection from images has a long and rich history of research in which discriminatively-trained scanning window classifiers [23]–[25] have proven to be both computationally efficient and accurate, the Viola Jones detector [12] is specifically well known due to it's implementation in a number of computer vision libraries. This method provides real-time face detection, but works best for full, frontal, and well

lit faces. In recent research deep CNN's have been applied which have shown state-of-the-art results specifically when dealing with non-frontal faces [14]–[16]. For a full review of face detection methods we refer the reader to the following survey papers [26]–[28].

B. LANDMARK LOCALISATION

Landmark localisation techniques have the aim of locating a set of facial landmarks, a large body of research has been conducted on this topic. Traditional approaches that have been applied within research includes AAM's [17] which apply a joint model of appearance and shape. Numerous improvements to the original AMM's have been proposed over time that improved the accuracy [29], [30]. Constrained Local Model [19], [31], [32] based techniques apply a local expert which provides a response filter which are constrained by a shape model. Regression based methods then gained popularity [20], [21], these techniques apply a set of features through regression based methods, which maps the discriminative features around landmarks to the desired landmark positions. Cascaded regression [33], [34] cascades a list of weak regressors to reduce the alignment error progressively. Initially deep learning based landmark localisation used regression based approaches applying CNN's to regress landmark locations, only a sparse set of landmarks where predicted in [35]–[37] unlike the 68 landmarks often used within traditional methods. In [38], convolutional layer outputs at different network levels were concatenated to predict 68 landmark locations. Following the introduction of the fully-convolutional network (FCN) [36] which produces facial landmark response maps with spatial equivalence to the raw images input, techniques based upon this idea became popular. Convolutional and de-convolutional networks where employed to generate the response map for each facial landmark, further localisation refinement was then applied utilising a network that performs regression in [39]–[41]. The stacked hourglass model was proposed in [42] for the task of human pose estimation which applied repeated bottom-up then top-down processing with intermediate supervision and obtained state-of-the-art result. This model has since been applied to the landmark localisation in [43] which proposed a binarized hourglass network for reduced computational complexity and [22] for 3D landmark generation. Dense 3D face alignment techniques such as the 3D morphable model (3DMM) [44], which aims to fit a 3D face shape to a 2D image also has the potential to deal with large poses. The 3D face shape is modelled using a linear subspace such as Tensor [45] or PCA [46] and achieves fitting through the minimisation of the difference between the model appearance and the image. These techniques can also suffer from a high computational cost though recently regression based 3DMM fitting [45], [47], [48] have improved on the efficiency.

C. INTEGRATED APPROACHES

One of the initial approaches to integrate both detecting faces and the associated landmarks was the tree structured model (TSM) [9] derived from the Deformable Part Model [13]. Not only did this model integrate tasks it also was conceived as a multi-view approach to deal with large pose variations. One considerable drawback of this method is the computational cost of deploying multiple models (up to 13 models) for the pose variation. Regressive Tree Structured Modal (RTSM) [49] is one method which applied a coarse-to-fine framework to reduce the computational overhead of the original TSM method. More recently [38] produced a method which uses a Fast R-CNN network and convolutional layer fusion for multitasks including face detection, landmark localisation and gender recognition. Reference [50] proposed a coarse-to-fine pipeline, face proposals are generated by a small fully convolutional network on the image pyramid. Face boxes are then classified and regressed to predict the five coarse facial landmarks. A similarity transformation is applied and the response map for each landmark estimate is calculated by the joint multi-view hourglass model.

III. METHOD

The proposed Integrated Deep Model for both face detection and landmark localisation from ''In The Wild'' images integrates two state-of-the-art architectures namely the Faster R-CNN network architecture [51] and the stacked hourglass based Face Alignment Network as defined in [22] respectively. The Faster R-CNN is traditionally applied to object detection tasks, we modify this and train the model solely for the task of face detection. The Face Alignment Network (FAN) [22] has previously been used for 68 point landmark localisation with an independent face detector in a linear based framework. While face detection recall is high in modern deep learning based techniques there is still room for further precision, this is specifically true when attempting to reduce the detection of false positives [52]–[54]. Our primary aim is to exploit the learnt features in both networks to provide a more precise face detection method while still providing accurate landmark localisation while maintaining computationally efficiency. Within this section we initially give an overview of the independent network architectures, then describe our novel Integrated Deep Model method. Fig. 2 provides a visual overview of our proposed IDM.

A. FASTER R-CNN FOR FACE DETECTION

A single deep CNN based upon the Faster R-CNN network architecture [51] has been shown to be a fast and accurate method for object detection, within this work we propose a variation we name Faster Face that is trained specifically for the task of face detection. The network architecture consists of three sections, firstly a VGG16 [2] architecture with five convolutional layers is applied which learns the features associated with the face detection task. The second is the region proposal network (RPN) layer which learns n regions of interests of probable face locations within images primarily for face detection. Finally a region of interest (ROI) pooling layer and a set of fully connected layers which correlate to the face detection prediction are present. The input to the

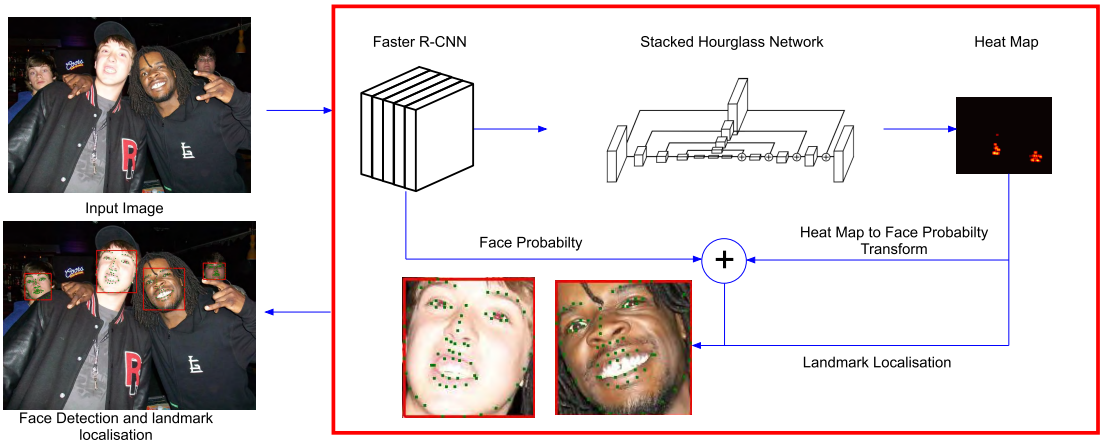

FIGURE 2. **The integrated deep model (IDM) a step-by-step overview. Note the FAN component is depicted as a single hourglass network when in reality it has 4 stacked together.**

proposed network is an N x M x 3 image, which are scaled and padded prior to entering the network if the native size is not N by M.

1) MULTI-TASK LOSS

The Faster R-CNN network [51] introduces a multi-task loss function to train the RPN and face detection tasks in a single CNN network. The total loss of the network is described as in (1).

$$loss_{total} = \sum_{i=n} loss_i \lambda_i \tag{1}$$

Total network loss as described in (1) where the individual loss corresponding to the ith task is defined as $loss_i$. A weighting parameter λ_i is applied to balance the learning priorities.

RPN Loss: The purpose of the RPN loss function is to learn regions based upon a set of potential anchor points of an image that most likely contain the desired object for detection which in this situation is a face. An anchor is box centred at a specific sliding window in the image input space, and is associated with a scale and aspect ratio. 3 scales and 3 aspect ratios are applied giving $n = 9$ anchors at each sliding window position. Given a convolutional feature map of size H x W derived from the N x M network input, total anchors $k = HWn$ where in this work $H = 39$ and $W = 51$. Given k anchors we assign a binary class label based upon the Intersection-over-Union (IoU). An anchor that has an IoU overlap higher than 0.7 is assigned a positive detection label while those anchors registering an IoU of less than 0.3 are labelled as negative. Other anchors which IoU value between 0.7 and 0.3 are not used for training.

$$loss_{cls} = \frac{1}{N_{cls}} \sum_{i=n} -(1-p_i^*) \cdot log(1-p_i) - p_i^* \cdot log(p_i) \tag{2}$$

$$loss_{reg} = \frac{1}{N_{reg}} \sum_{i=n} p_i^* smooth_{L1}(t_i - t_i^*) \tag{3}$$

The softmax loss function given by (2) is used for learning an object ($p_i = 1$) and a non-object ($p_i = 0$) classification, where p_i^* is the ground truth class label and p_i the predicted class for the ith anchor respectively. This loss function is normalised by N_{cls} which is the mini-batch size. Bounding box regression is defined as (3), where for the ith anchor the $L1$ loss between the ground-truth box t_i^* and the predicted bounding box t_i is calculated. Both t_i^* and t_i are vectors representing the 4 parameterised coordinates of the predicted bounding box. Only positive anchors affect the loss as described by the term $p_i^* smooth_{L1}$. N_{reg} the total number of anchors normalises the loss function. For a full technical overview of the RPN architecture we refer the reader to [51].

Face Detection Loss: Our proposed network redefines the object detection task of [51] for the purpose of face detection. Face detection is a binary class problem of a face or not within a proposed region, and the regression of a bounding box for the location of the face within an image. Face detection loss applies similar loss functions to the RPN loss as defined in (2) and (3). Principally the difference between RPN loss and face detection loss is that RPN loss is concerned with finding a subset of anchors which best describe objects, where as face detection loss learns whether these anchors contain a face or not.

B. FACE ALIGNMENT NETWORK - LANDMARK LOCALISATION

The state-of-the-art stacked hourglass architectures have been shown to be highly accurate for tasks including human pose analysis and facial landmark localisation. Within this paper we specifically implement the model as defined in [22] to provide 68 localised 2D facial landmarks for the detected face. The hourglass network was initially proposed in [42] and takes it's name from the construction of the layers of the network which can be seen in Fig. 2. Initially convolutional and max pooling layers are used to process features down to a very low resolution, during this down sampling of the input the network branches off prior to each max pooling step and further convolutions are applied on the pre-pooled branches, this is then fed back into the network during up sampling.

FIGURE 3. Example of false positive detection's. (A) - Faces unlabelled in the ground truth data (B) - Tiny detection's from the faster face (C) - The remaining detection's.

Following the lowest level of convolution the network then begins to up sample back to the original resolution through the application of nearest neighbour up sampling and element wise addition of the previously branched features. The implemented method stacks four hourglass networks and for each convolutional layer uses the hierarchical, parallel and multi-scale block [43], which performs three levels of parallel convolution alongside batch normalisation before outputting the concatenated feature map. The output of this network is a set of heat maps where for a given heat map the network predicts the probability of a facial landmarks presence at each and every pixel of the inputted image.

C. INTEGRATED DEEP MODEL

Our hypothesis is that the features learnt for landmark localisation have inter-connectivity with the task of face detection. Given the two independent architectures discussed previously in this section we now describe our proposed novel method to combine the networks creating our Integrated Deep Model. To achieve this integration we define the following heat map transformation, integrated loss function using a joint probability for face detection and size scaling techniques while adding minimal computational overhead when compared to using the two architectures in a linear framework.

Given the heat map output of the FAN as $H = h_1, h_2, ..., h_n$ where each h_i is a n x m matrix equal in dimensions to the input image for the i^{th} facial landmark. Each value in h_i corresponds to the probability of the facial landmark being located at that specific pixel location within a given input image. We propose a method as described within (4) to transform the heat map H to a probability score that can be applied to the task of face detection and integrate this with the loss function of the Faster R-CNN face detector.

$$p_{fan} = \frac{1}{N} \sum_{i=n} \max(H_i)\gamma_i \tag{4}$$

Given by the maximum probability $\max(H_i)$ for the i^{th} facial landmark a specific scaling factor γ_i is applied for that that landmark. The sum of the scaled probability is then normalised and can be considered as the probability of a face detection derived from the FAN network defined as p_{fan}. The scaling value γ is primarily introduced to deal with wide ranging face poses in which certain landmarks retain visibility across all poses where others become occluded, the values of γ_i used are reported within our intermediate results in section IV.

$$p_{face} = \frac{(p_{fan} + (p_{faster}\delta))}{2} \tag{5}$$

The next step is to define the joint probability of a region being a face as termed as p_{face} and defined in (5) where p_{faster} is the probability based upon the output of the trained Faster Face features. The penalisation factor δ is specifically introduced for situations where extremely small detection's are classed in the very high 90% probability range as being faces when they are not (Fig. 3 provides examples of this). The value of δ is determined by (6) where det is the width of the face detection box and img is the width of the image,

we only apply probability penalisation when a face width is less that 2% of the total image width. Finally the p_{face} is used within the loss function for the face detection classification as described in (7).

$$\delta = \begin{cases} 0.7 & \text{if det * (100 / img)} \leq 2 \\ 1 & \text{otherwise} \end{cases} \tag{6}$$

$$loss_{face} = \frac{1}{N_{cls}} \sum_{i=n} -(1 - p_i^*) \cdot log(1 - p_{face,i}) - p_i^* \cdot log(p_{face,i}) \tag{7}$$

D. MODEL TRAINING

Our IDM method specifically trains the layers of a Faster R-CNN architecture [4] with images containing multiple faces from the Wider Face training set [55]. The popular augmentation method of flipping is employed to further the available training data. Our faster R-CNN implementation applies a VGG16 [2] architecture for the convolutional layers and the weight parameters are initiated using a pre-trained imagenet model prior to fine tuning for face detection. The publicly available PyTorch pre-trained model of the Facial Alignment Network of [22] is used as the stacked hourglass component of IDM. Our IDM model was trained for 15 epochs with a learning rate of 10^{-3} for the initial 10 epochs then 10^{-4} for the final 5.

IV. EXPERIMENTAL EVALUATION

Within this section we present a thorough experimental evaluation of our proposed IDM method in the areas of face detection and landmark localisation. All experiments are conducted using PyTorch 0.4 on Windows 10 with a Nvidia GTX 1080 GPU.

A. FACE DETECTION EVALUATION

To evaluate the effectiveness of our proposed IDM method for the task of face detection we produce a set of intermediate and benchmark evaluations. For robust evaluation we apply three face detection test sets, these being the Annotated Faces In-The-Wild (AFW) database [9], the Annotated Facial Landmarks In The Wild (AFLW) [56] and the Face Detection Data set and Benchmark (FDDB) [57]. The AFW database consists of 205 images where each image contains at least a single face, in total there are 468 faces located within the database. The AFLW database test set contains around 1,001 images of annotated faces in real-world images capturing multiple viewpoints, different expressions and illumination conditions. Finally the FDDB consists of 5171 faces in 2845 images from unconstrained environments. We adopt the PASCAL VOC precision-recall protocol for object detection requiring 50% IoU for positive detection of a face.

1) INTERMEDIATE RESULTS

For our intermediate results four different methods are evaluated, these being three variations of the IDM method using different parameters and a standalone Faster R-CNN only face detector we call Faster Face. The three IDM variations are as follows, IDM Mean where the $\gamma = 1$ for all 68 landmarks, IDM Scaled which applies varying values for γ, the landmarks that are visible across all facial poses such as the nose are given a value of 1 while other less visible landmarks are given γ values of 0.75. Finally IDM Scale and Sized add a box size weighting factor penalty as described within the previous section.

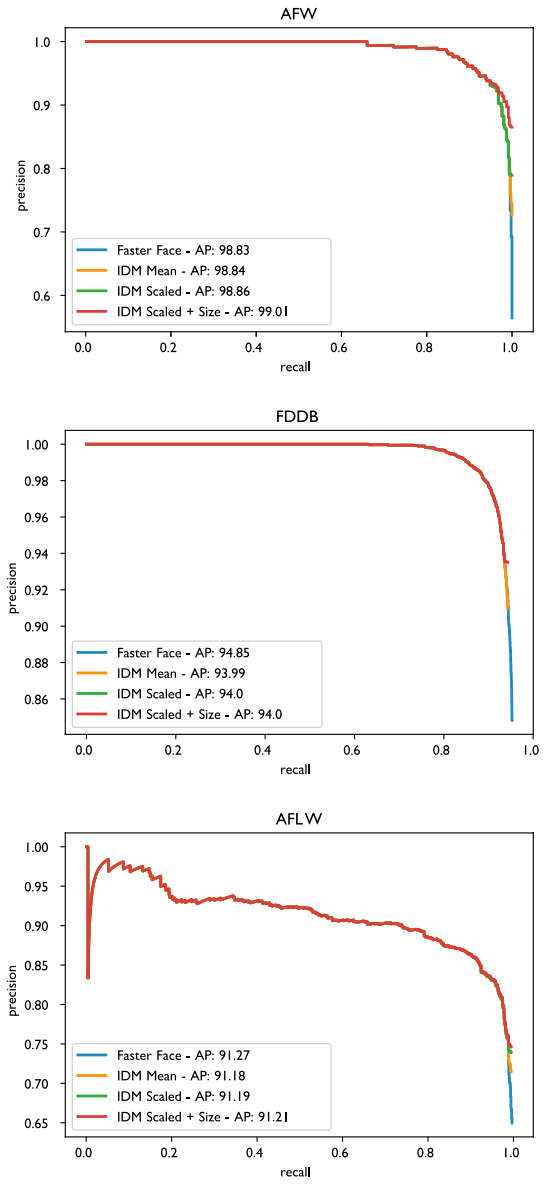

FIGURE 4. Intermediate results precision-recall curves.

The results for AFW test set are given in Fig. 4 and Table 1, we show 100% recall of the faces for all methods. While all IDM methods dramatically reduce the number of false positives, the greatest reduction being from 361 to 73 which is significant. The IDM Scaled and Size also has success in correctly identifying the small detection's (examples of this are shown within Fig. 3). For the FDDB test set the results are

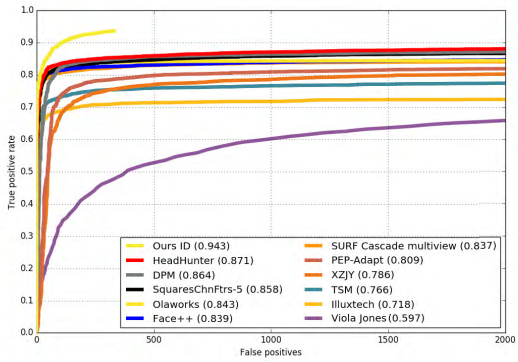

FIGURE 5. FDDB benchmark results.

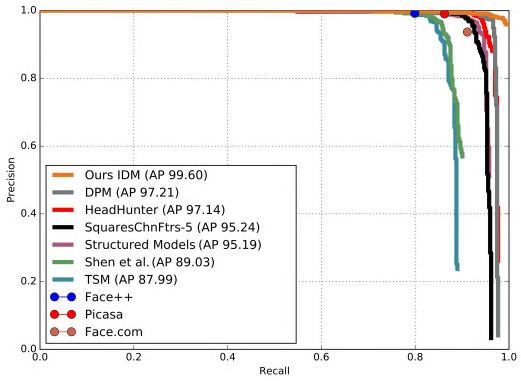

FIGURE 6. AFW benchmark results.

TABLE 1. AFW results benchmark.

Method	True Positives	False Positives
Faster Face	468	361
IDM Mean	468	176
IDM Scaled	468	125
IDM Scaled + Size	468	73

TABLE 2. FDDB results benchmark.

Method	True Positives	False Positives
Faster Face	4926	882
IDM Mean	4876	484
IDM Scaled	4876	336
IDM Scaled + Size	4876	336

highlighted in Fig. 4 and Table 2, again a large reduction of over 50% is shown in the false positives for IDM methods. A small decrease in recall is noticed in comparison with Faster Face, when analysed this drop is almost entirely for very blurred faces (see Fig. 7). Finally AFLW test set results are shown in Fig. 4 and Table 3, this follows a similar pattern to the previous results with a minimal drop in recall of 2 faces, but a significant drop of 286 false positives between the Faster Face method and the full IDM method.

The overall observations from all three test sets is that our proposed IDM method and most specifically the IDM Scale and Sized variation has a large impact on reducing false

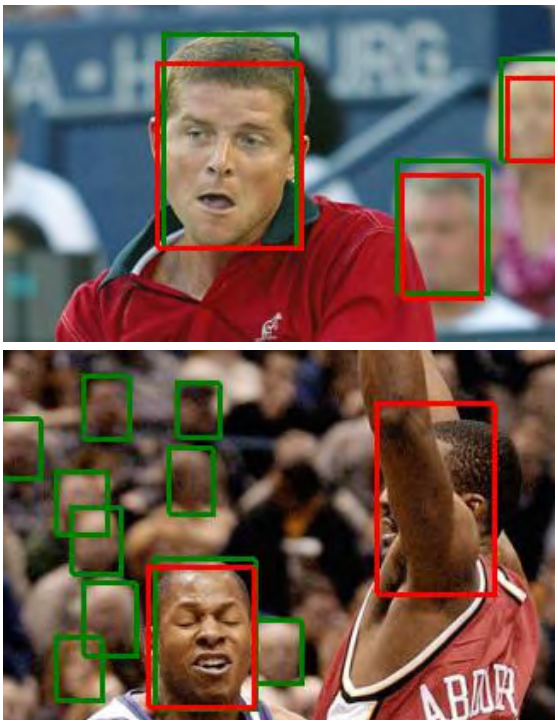

FIGURE 7. Blurred faces detection's examples (Red boxes represent a detection by IDM, green boxes are the ground truth face data). The top image displays example of moderately blurred faces which our models successfully detects. The bottom images highlights extreme blur where the IDM method misses 9 faces, while also detecting a face not accounted for in the ground truth data.

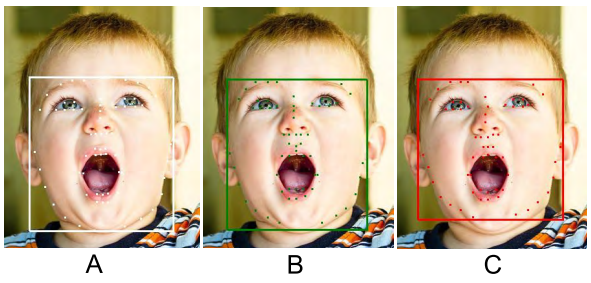

FIGURE 8. Example of Face Bounding Box affecting landmark localisation accuracy. (A) - Ground truth data (B) - FAN using ground truth bounding box (C) - IDM (Predicted bounding box does not cover the point of the chin which affects the landmark accuracy around the jaw line).

TABLE 3. AFLW results benchmark.

Method	True Positives	False Positives
Faster Face	1183	638
IDM Mean	1181	471
IDM Scaled	1181	368
IDM Scaled + Size	1181	352

positive. The negative aspects is that there is a small reduction in recall from analysis of the images we find that the primary source of decreased recall is blurred faces. The results also highlights the effectiveness of the Faster Face architecture alone for detecting faces within "In The Wild" images in terms of recall, where this has issue is that it also provides a high amount of false positives.

FIGURE 9. Example of false positive detection's (red boxes represent a detection by the IDM, green boxes are the ground truth face data). (A) - Faces unlabelled in the ground truth data (B) - Overlapping boxes on a single face (C) - Detected face box to small in comparison to eh ground truth to be within the metrics applied (D) - Non face detection's.

2) BENCHMARK RESULTS

To benchmark against other face detection methods we use the AFW and FDDB test sets, as the AFLW test set is not a standardised image set. The evaluation tool provided by [52] is used for results generation and provides benchmarks against a number of methods including Headhunter [52], Structured Models [58] and Tree Shape Models [9]. Note this software uses an alternative set of ground truth boxes for the AFW explaining the difference in average precision compared with our previous intermediate evaluation. For the FDDB and AFW test set our method outperforms the methods included within the evaluation tool, we show both higher recall and also less false positives by a significant margin in both cases as shown in Fig. 5 and Fig. 6. Furthermore when we compare our method to other recent state-of-the-art CNN based methods the Fast Deep Convolutional (FD-CNN) [54] and the impressive S^3FD [59] we have very competitive results in terms of recall while having much lower rates of false positive detection's. For the FDDB test set FD-CCN has a recall rate of 92.6% while S^3FD has a 98.2%, our method is in between at 94.3%. Where our method excels is in the false positives which are significantly lower than both methods at 336 compared to 700 and over 1000 for FD-CNN and S^3FD respectively. Only the S^3FD provides benchmarks for the AFW in which they report at 99.8% average precision compared to our 99.6%, again our IDM method has greater precision [59]. As identified in our intermediate results the main recall issue for our model is severely blurred faces (see Fig. 7 for examples), primarily due to the training set not containing blurred faces to this degree.

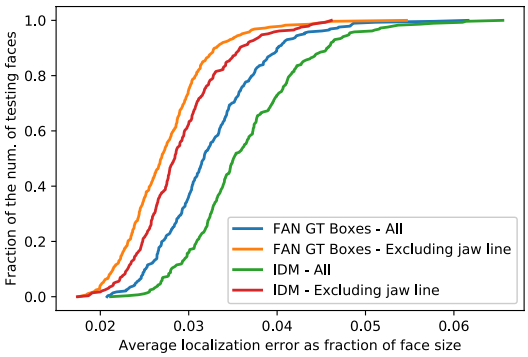

FIGURE 10. Cumulative localisation error distribution from 300W test set.

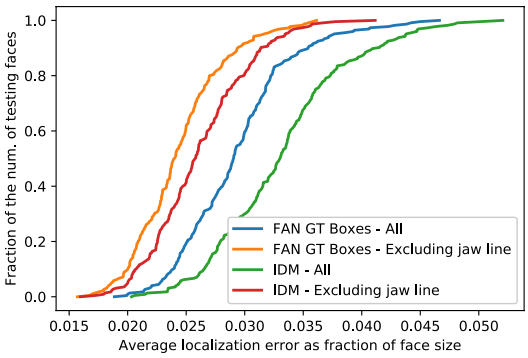

FIGURE 11. Cumulative localisation error distribution from AFW test set.

B. LANDMARK LOCALISATION EVALUATION

The primary objective for our proposed method is to improve the accuracy of face detection which we have analysed in

the previous experiments, as the landmark localisation features of the method are not re-trained we do not expect to observe increased accuracy over the baseline presented in [22]. Instead we evaluate the performance of face alignment accuracy in IDM against the base line experiments for the purpose of understanding how face bounding box affects landmark localisation accuracy and to what degree. This mimics a more real world application of the landmark localisation where ground truth face bounding boxes are not provided. We evaluate the landmark localisation predicted by the proposed IDM on the 300-W test set [60], which consists of 600 fully annotated faces and the AFW test set previously used in the face detection evaluation using the IBUG annotations. Normalised Mean Error (NME) using face size normalisation as described in [22] is used as the evaluation metric. Cumulative localisation error is shown in Fig. 10 and Fig. 11. For both evaluations the results are similar, there is a slightly larger error margin when using the bounding boxes from our IDM method compared to the using ground truth boxes. This outcome is not surprising but highlight's the importance of bounding box accuracy even with state-of-the-art landmark localisation techniques. Accuracy with the IDM method is in general high at its largest we observe around a 0.005 difference in error as a fraction of the face size. This suggests the IDM method has good bounding box accuracy and the landmark localisation is somewhat robust to initialisation. The largest error is for those landmarks that make up the jaw line which seem to be the most influenced by bounding box placement. One reason for this is in cases where the predicted face detection bounding box does not cover the entire face. An example of the effect of bounding box on landmark localisation can be found in Fig. 8.

V. CONCLUSIONS

Within this paper we present our proposed Integrated Deep Model for detecting faces and performing landmark localisation from unconstrained "In The Wild" images. Our primary contribution is to produce a optimisation method to successfully integrate features from two state-of-the-art deep learning architectures to leverage both their strengths for more precise face detection. We show that our method is comparable to other top performing face detection methods on the AFW and FDDB test sets. Specifically IDM show very significant reductions on the number of false positive detection's increasing the precision while having a very small impact on recall. The analysis of landmark localisation performed by the IDM on the 300-W and AFW test sets highlights high accuracy though when compared with landmark localisation using ground truth face bounding boxes there is a small increase in error specifically for the landmarks which border the face such as the jaw line. The main cause of error is the precision of the face bounding box.

There are a number of areas to be considered for further research, while the IDM method reduces the number of false positives dramatically some still remain, from analysis of the data these can be labelled into four distinct categories and are shown in Fig. 9. The first category is that of faces in images that are missed from the ground truth labelling of the test data, the second are multiple boxes that are overlapping the same face but are not removed by Non-maximal suppression. Thirdly are faces that are found but the detected bounding box is not precise enough to be classified as a true detection, this is most common in occluded faces. Finally we simply have wrong detection's. Further investigating methods to better predict face bounding boxes potentially with techniques such as edge detection or image segmentation could lead to better face detection accuracy and this in turn would also help to increase landmark localisation accuracy.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[2] K. Simonyan and A. Zisserman. (Sep. 2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[5] J. Redmon and A. Farhadi. (Dec. 2016). "YOLO9000: Better, faster, stronger." [Online]. Available: https://arxiv.org/abs/1612.08242

[6] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Toward a practical face recognition system: Robust alignment and illumination by sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 372–386, Feb. 2012.

[7] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[8] X. Li *et al.* (Nov. 2015). "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods." [Online]. Available: https://arxiv.org/abs/1511.00423

[9] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2879–2886.

[10] H. Yang, W. Mou, Y. Zhang, I. Patras, H. Gunes, and P. Robinson. (Jul. 2015). "Face alignment assisted by head pose estimation." [Online]. Available: https://arxiv.org/abs/1507.03148

[11] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published. [Online]. Available: https://ieeexplore.ieee.org/document/8170321

[12] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2009.

[14] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.

[15] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2016, vols. 11–18, no. 3, pp. 3676–3684.

[16] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in *Proc. Int. Conf. Multimedia Retr. (ICMR)*, 2015, pp. 643–650.

[17] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[18] R. Gross, I. Matthews, and S. Baker, "Active appearance models with occlusion," *Image Vis. Comput.*, vol. 24, no. 6, pp. 593–604, 2006.

[19] D. Cristinacce and T. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognit.*, vol. 41, no. 10, pp. 3054–3067, 2008.

[20] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2729–2736.

[21] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 FPS via regressing local binary features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1685–1692.

[22] A. Bulat and G. Tzimiropoulos. (Sep. 2017). "How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks)." [Online]. Available: https://arxiv.org/abs/1703.07332

[23] M. Jones and P. Viola, "Fast multi-view face detection," Mitsubishi Electr. Res. Lab., Cambridge, MA, USA, Tech. Rep. TR20003-96, Jul. 2003.

[24] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[25] B. Heisele, T. Serre, and T. Poggio, "A component-based framework for face detection and identification," *Int. J. Comput. Vis.*, vol. 74, no. 2, pp. 167–181, 2007.

[26] M.-H. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 1, pp. 34–58, Jan. 2002.

[27] C. Zhang and Z. Zhang, "A survey of recent advances in face detection," Microsoft Res., Cambridge, U.K., Tech. Rep. MSR-TR-2010-66, Jun. 2010.

[28] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: Past, present and future," *Comput. Vis. Image Understand.*, vol. 138, pp. 1–24, Sep. 2015.

[29] J. Saragih and R. Goecke, "A nonlinear discriminative approach to AAM fitting," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[30] G. Tzimiropoulos and M. Pantic, "Optimization problems for fast AAM fitting in-the-wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 593–600.

[31] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *Int. J. Comput. Vis.*, vol. 91, no. 2, pp. 200–215, Jan. 2011.

[32] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3444–3451.

[33] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1078–1085.

[34] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4998–5006.

[35] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3476–3483.

[36] Z. Liang, S. Ding, and L. Lin. (Jul. 2015). "Unconstrained facial landmark localization with backbone-branches fully-convolutional networks." [Online]. Available: https://arxiv.org/abs/1507.03409

[37] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Computer Vision—ECCV* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8694. Cham, Switzerland: Springer, 2014, pp. 94–108. [Online]. Available: https://link.springer.com/book/10.1007/978-3-319-10599-4#volumes

[38] R. Ranjan, V. M. Patel, and R. Chellappa. (Mar. 2016). "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition." [Online]. Available: https://arxiv.org/abs/1603.01249

[39] H. Lai *et al.* (Oct. 2015). "Deep recurrent regression for facial landmark detection." [Online]. Available: https://arxiv.org/abs/1510.09083

[40] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim, "Robust facial landmark detection via recurrent attentive-refinement networks," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2016, pp. 57–72.

[41] A. Bulat and Y. Tzimiropoulos, "Convolutional aggregation of local evidence for large pose face alignment," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 1–86.

[42] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2016, pp. 483–499.

[43] A. Bulat and G. Tzimiropoulos. (Aug. 2017). "Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources." [Online]. Available: https://arxiv.org/abs/1703.00862

[44] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. 26th Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*. New York, NY, USA: ACM Press, 1999, pp. 187–194.

[45] C. Cao, Q. Hou, and K. Zhou, "Displaced dynamic expression regression for real-time facial tracking and animation," *ACM Trans. Graph.*, vol. 33, no. 4, 2014, Art. no. 43.

[46] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1063–1074, Sep. 2003.

[47] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas, "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 1944–1951.

[48] L. A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3D face alignment from 2D videos in real-time," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 1, May 2015, pp. 1–8.

[49] G.-S. Hsu, K.-H. Chang, and S.-C. Huang, "Regressive tree structured model for facial landmark localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3855–3861.

[50] J. Deng, G. Trigeorgis, Y. Zhou, and S. Zafeiriou. (Aug. 2017). "Joint multi-view face alignment in the wild." [Online]. Available: https://arxiv.org/abs/1708.06023

[51] S. Ren, K. He, R. Girshick, and J. Sun. (Jun. 2015). "Faster R-CNN: Towards real-time object detection with region proposal networks." [Online]. Available: https://arxiv.org/abs/1506.01497

[52] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2014, pp. 720–735.

[53] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. (Nov. 2017). "S^3 FD: Single shot scale-invariant face detector." [Online]. Available: https://arxiv.org/abs/1708.05237

[54] D. Triantafyllidou, P. Nousi, and A. Tefas, "Fast deep convolutional face detection in the wild exploiting hard sample mining," *Big Data Res.*, vol. 11, pp. 65–76, Mar. 2018.

[55] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Dec. 2016, pp. 5525–5533.

[56] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2144–2151.

[57] V. Jain, V. Jain, and E. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings," Dept. Comput. Sci., Univ. Massachusetts, Boston, MA, USA, Tech. Rep. UM-CS-2010-009, 2010.

[58] J. Yan, X. Zhang, Z. Lei, and S. Z. Li, "Face detection by structural models," *Image Vis. Comput.*, vol. 32, no. 10, pp. 790–799, 2014.

[59] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. (Nov. 2017). "S^3 FD: Single shot scale-invariant face detector." [Online]. Available: https://arxiv.org/abs/1708.05237

[60] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 397–403.

GARY STOREY received the M.Sc. degree in computer science from Northumbria University, Newcastle upon Tyne, U.K., in 2015, where he is currently pursuing the Ph.D. degree with the Department of Computer and Information Sciences.

His research interests include computer vision, machine learning, and human face analysis and the potential application domains. He received the Best Student Paper from the Intelligent Systems Conference (IntelliSys) 2018.

AHMED BOURIDANE received the Ingenieur d'État degree in electronics from the Ecole Nationale Polytechnque of Algiers, Algeria, in 1982, the M.Phil. degree in electrical engineering (VLSI design for signal processing) from the University of Newcastle, Newcastle Upon Tyne, U.K., in 1988, and the Ph.D. degree in electrical engineering (computer vision) from the University of Nottingham, U.K., in 1992.

From 1992 to 1994, he was a Research Developer in telesurveillance and access control applications. In 1994, he joined Queen's University Belfast, Belfast, U.K., initially as a Lecturer in computer architecture and image processing and then as a Reader in computer science. He became a Professor in image engineering and security at Northumbria University, Newcastle, U.K., in 2009. He has authored and co-authored more than 200 publications. His research interests are in imaging for forensics and security, biometrics, homeland security, image/video watermarking, and cryptography.

RICHARD JIANG received the Ph.D. degree in computer science from Queen's University Belfast, Belfast, U.K., in 2008. He is currently a Senior Lecturer with the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne, U.K.

From 2007 to 2013, he was with Brunel University London, Loughborough University, Swansea University, University of Bath, and The University of Sheffield. He joined Northumbria University in 2013. He has authored and co-authored more than 50 publications. His research interests mainly reside in the fields of artificial intelligence, biometrics, privacy and security, and biomedical image analysis. His research has been funded by EPSRC, BBSRC, TSB, EU FP, and industry funds.